IRASubcat, a highly customizable, language independent tool for the acquisition of verbal subcategorization information from corpus

Ivana Romina Altamirano and **Laura Alonso i Alemany**
Grupo de Procesamiento de Lenguaje Natural
Sección de Ciencias de la Computación
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina
romina.altamirano@gmail.com,alemany@famaf.unc.edu.ar

Abstract

IRASubcat is a language-independent tool to acquire information about the subcategorization of verbs from corpus. The tool can extract information from corpora annotated at various levels, including almost raw text, where only verbs are identified. It can also aggregate information from a pre-existing lexicon with verbal subcategorization information. The system is highly customizable, and works with XML as input and output format.

IRASubcat identifies patterns of constituents in the corpus, and associates patterns with verbs if their association strength is over a frequency threshold and passes the likelihood ratio hypothesis test. It also implements a procedure to identify verbal constituents that could be playing the role of an adjunct in a pattern. Thresholds controlling frequency and identification of adjuncts can be customized by the user, or else they are given a default value.

1 Introduction and Motivation

Characterizing the behavior of verbs as nuclear organizers of clauses (the so-called subcategorization information) is crucial to obtain deep analyses of natural language. For example, it can significantly reduce structural ambiguities in parsing (Carroll et al., 1999; Carroll and Fang, 2004), help in word sense disambiguation or improve information extraction (Surdeanu et al., 2003). However, the usual construction of linguistic resources for verbal subcategorization involves many expert hours, and it is usually prone to low coverage and inconsistencies across human experts.

Corpora can be very useful to alleviate the problems of low coverage and inconsistencies. Verbs can be characterized by their behavior in a big corpus of the language. Thus, lexicographers only need to validate, correct or complete this digested information about the behavior of verbs. Moreover, the starting information can have higher coverage and be more unbiased than if it is manually constructed. That's why automatic acquisition of subcategorization frames has been an active research area since the mid-90s (Manning, 1993; Brent, 1993; Briscoe and Carroll, 1997).

However, most of the approaches have been adhoc for particular languages or particular settings, like a determined corpus with a given kind of annotation, be it manual or automatic. To our knowledge, there is no system to acquire subcategorization information from corpora that is flexible enough to work with different languages and levels of annotation of corpora.

We present IRASubcat, a tool that acquires information about the behaviour of verbs from corpora. It is aimed to address a variety of situations and needs, ranging from rich annotated corpora to virtually raw text (because the tags to study can be selected in the configuration file). The characterization of linguistic patterns associated to verbs will be correspondingly rich. The tool allows to customize most of the aspects of its functioning, to adapt to different requirements of the users. Moreover, IRASubcat is platform-independent and open source, available for download at `http://www.irasubcat.com.ar`.

IRASubcat input is a corpus (in xml format) with examples of the verbs one wants to characterize, and its output is a lexicon where each verb is associated with the patterns of linguistic constituents that reflect its behavior in the given corpus, an approxima-

tion to its subcategorization frame. Such association is established when the verb and pattern co-occur in corpus significantly enough to pass a frequency test and a hypothesis test.

In the following section we discuss some previous work in the area of subcategorization acquisition from corpora. Then, Section 3 presents the main functionality of the tool, and describe its usage. Section 4 details the parameters that can be customized to adapt to different experimental settings. In Section 5 we outline the functionality that identifies constituents that are likely to be adjuncts and not arguments, and in Section 6 we describe the procedures to determine whether a given pattern is actually part of the subcategorization frame of a verb. Section 7 presents some results of applying IRASubcat to two very different corpora. Finally, we present some conclusions and the lines of future work.

2 Previous Work

We review here some previous work related to acquisition of subcategorization information from corpora, focussing on the constraints of the approach and corpora to learn with. We specially mention approaches for languages other than English.

The foundational work of (Brent, 1993) was based on plain text (2.6 million words of the Wall Street Journal (WSJ, 1994)). Since the corpus had no annotation, verbs were found by heuristics. He detected six frame types and filtered associations between verbs and frames with the binomial hypothesis test. This approach obtained 73.85% f-score in an evaluation with human judges.

Also in 1993, (Ushioda et al., 1993) exploited also the WSJ corpus but only the part that was annotated with part-of-speech tags, with 600.000 words. He studied also six frame types and did not distinguishing arguments and adjuncts.

The same year, (Manning, 1993) used 4 million words of the New York Times (Sandhaus,), selected only clauses with auxiliary verbs and automatically analyzed them with a finite-state parser. He defined 19 frame types, and reported an f-score of 58.20%.

Various authors developed approaches assuming a full syntactic analysis, which was usually annotated manually in corpora (Briscoe and Carroll, 1997; Kinyon and Prolo, 2002). Others associated syntactic analyses to corpora with automatic parsers (O'Donovan et al., 2005).

Various approaches were also found for languages other than English. For German, (Eckle-Kohler, 1999) studied the behaviour of 6305 verbs on automatically POS-tagged corpus data. He defined linguistic heuristics by regular expression queries over the usage of 244 frame types.

(Wauschkuhn, 1999) studied 1044 German verbs. He extracted maximum of 2000 example sentences for each verb from a corpus, and analyzed them with partial (as opposed to full) syntactic analysis. He found valency patterns, which were grouped in order to extract the most frequent pattern combinations, resulting in a verb-frame lexicon with 42 frame types.

(Schulte im Walde, 2000) worked with 18.7 million words of German corpus, found 38 frame types. She used the Duden das Stilwörterbuch(AG, 2001) to evaluate results and reported f-score 57,24% with PP and 62,30% without.

Many other approaches have been pursued for various languages: (de Lima, 2002) for Portuguese, (Georgala, 2003) for Greek, (Sarkar and Zeman, 2000) for Czech, (Spranger and Heid, 2003) for Dutch, (Chesley and Salmon-Alt, 2006) for French or (Chrupala, 2003) for Spanish, to name a few.

3 General description of the tool

IRASubcat takes as input a corpus in XML format. This corpus is expected to have some kind of annotation associated to its elements, which will enrich the description of the patterns associated to verbs. The minimal required annotation is that verbs are marked. If no other information is available, the form of words will be used to build the patterns. If the corpus has rich annotation for its elements, the system can build the patterns with the value of attributes or with a combination of them, and also with combinations with lexical items. The only requirements are that verbs are marked, and that all linguistic units to be considered to build the patterns are siblings in the XML tree.

The output of IRASubcat is a lexicon, also in XML format, where each of the verbs under inspection is associated to a set of subcategorization patterns. A given pattern is associated to a given verb

if the evidence found in the corpus passes certain tests. Thresholds for these tests are defined by the user, so that precision can be priorized over recall or the other way round. In all cases, information about the evidence found and the result of each test is provided, so that it can be easily assessed whether the threshold for each test has the expected effects, and it can be modified accordingly.

The lexicon also provides information about frequencies of occurrence for verbs, patterns, and their co-occurrences in corpus.

Moreover, IRASubcat is capable of integrating the output lexicon with a pre-existing one, merging information about verbs and patterns with information that had been previously extracted, possibly from a different corpus or even from a hand-built lexicon. The only requirement is that the lexicon is in the same format as IRASubcat output lexicon.

4 A highly customizable tool

IRASubcat has been designed to be adaptable in a variety of settings. The user can set the conditions for many aspects of the tool, in order to extract different kinds of information for different representational purposes or from corpora with different kinds of annotation. For example, the system accepts a wide range of levels of annotation in the input corpus, and it is language independent. To guarantee that any language can be dealt with, the corpus needs to be codified in UTF-8 format, in which virtually any existing natural language can be codified.

If the user does not know how to customize these parameters, she can resort to the default values that are automatically provided by the system for each of them. The only information that needs to be specified in any case is the name of the tag marking verbs, the name of the parent tag for the linguistic units that characterize patterns and, of course, the input corpus.

The parameters of the system are as follows:

- The user can provide a list of verbs to be described, so that any other verb will not be considered. If no list is provided, all words marked as verb in the corpus will be described.
- The scope of patterns can be specified as a window of n words around the words marked as verbs, where n is a number specified by the user. It can also be specified that all elements that are siblings of the verb in the XML tree are considered, which is equivalent to considering all elements in the scope of the clause, if that is the parent node of the verb in an annotated corpus. By default, a window of 3 sibling nodes at each side of the verb is considered.
- It can be specified that patterns are completed by a dummy symbol if the context of occurrence of the verb does not provide enough linguistic elements to fill the specified window length, for example, at the end of a sentence. By default, no dummy symbol is used.
- It can be specified whether the order of occurrence of linguistic units should be taken into account to characterize the pattern or not, depending of the meaning of word order in the language under study. By default, order is not considered.
- We can provide a list of the attributes of linguistic units that we want to study, for example, syntactic function, morphological category, etc. Attributes should be expressed as an XML attribute of the unit. It can also be specified that no attribute of the unit is considered, but only its content, which is usually the surface form of the unit. By default, an attribute named “sint” will be considered.
- We can specify whether the content of linguistic units will be considered to build patterns. As in the previous case, the content is usually the surface form of the unit (lexical form). By default, content is not considered.
- A mark describing the position of the verb can be introduced in patterns. By default it is not considered, to be coherent with the default option of ignoring word order.
- It can be specified that, after identifying possible adjuncts, patterns with the same arguments are collapsed into the same pattern, with all their characterizing features (number of occurrences, etc.). By default, patterns are not collapsed.
- The number of iterations that are carried out on patterns to identify adjuncts can be customized,

by default it is not considered because by default patterns are not collapsed.

- The user can specify a minimal number of occurrences of a verb to be described. By default, the minimal frequency is 0, so all verbs that occur in the corpus are described.
- A minimal number of occurrences of a pattern can also be specified, with the default as 0.
- The user can specify whether the Log-Likelihood Ratio hypothesis test will be applied to test whether the association between a verb and a pattern cannot be considered a product of chance. By defect, the test is used (and the output will be 90, 95, 99 or 99.5 when the co-ocurrence have that confiability) .

5 Identification of adjuncts

One of the most interesting capabilities of IRASubcat is the identification of possible adjuncts. Adjuncts are linguistic units that do not make part of the core of a subcategorization pattern (Fillmore, 1968). They are optional constituents in the constituent structure governed by a verb. Since they are optional, we assume they can be recognized because the same pattern can occur with or without them without a significant difference. IRASubcat implements a procedure to identify these units by their optionality, described in what follows. An example of this procedure is shown in Figure 1.

First, all patterns of a verb are represented in a trie. A trie is a tree-like structure where patterns are represented as paths in the trie. In our case, the root is empty and each node represents a constituent of a pattern, so that a pattern is represented by concatenating all nodes that are crossed when following a path from the root. Each node is associated with a number expressing the number of occurrences of the pattern that is constructed from the root to that node. Constituents are ordered by frequency, so that more frequent constituents are closer to the root.

In this structure, it is easy to identify constituents that are optional, because they are topologically located at the leaves of the trie and the number of occurrences of the optional node is much smaller than the number of occurrences of its immediately preceding node.

We have experimented with different ratios between the frequency of the pattern with and without the constituent to identify adjuncts. We have found that adjuncts are usually characterized by occurring in leaves of the trie at least for 80% of the patterns of the verb.

Once a constituent is identified as an adjunct, it is removed from all patterns that contain it within the verb that is being characterized at the moment. A new trie is built without the adjunct, and so new adjuncts may be identified. This procedure can be iterated until no constituent is found to be optional, or until a user-defined number of iterations is reached.

When an adjunct is removed, the original pattern is preserved, so that the user can see whether a given pattern occurred with constituents that have been classified as adjuncts, and precisely which constituents.

When this data structure is created, the sequential ordering of constituents is lost, in case it had been preserved in the starting patterns. If the mark signalling the position of the verb had been introduced, it is also lost. However, order and position of the verb can be recovered in the final patterns, after adjuncts have been identified.

6 Associating patterns to verbs

One of the critical aspects of subcategorization acquisition is the association of verbs and patterns. How often must a pattern occur with a verb to make part of the subcategorization frame of the verb? To deal with this problem, different approaches have been taken, going from simple co-occurrence count to various kinds of hypothesis testing (Korhonen et al., 2000).

To determine whether a verb and a pattern are associated, IRASubcat provides a co-occurrence frequency threshold, that can be tuned by the user, and a hypothesis test, the Likelihood Ratio test (Dunning, 1993). We chose to implement this test, and not others like the binomial that have been extensively used in subcategorization acquisition, because the Likelihood Ratio is specially good at modeling unfrequent events.

To perform this test, the null hypothesis is that the distribution of an observed pattern 'M_j' is independent of of the distribution of verb 'V_i'.

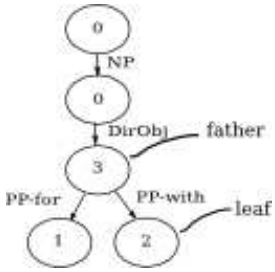
Figure 1: Example of application of the procedure to identify adjuncts.

1. A starting set of patterns:
   `[NP DirObj PP-with], [NP DirObj], [NP DirObj], [NP DirObj PP-with],[NP DirObj] y [NP DirObj PP-for]`
2. Pattern constituents are ordered by frequency:
   `NP > DirObj > PP-with > PP-for`
3. Constituents in patterns are ordered by their relative frequency:

   ```
   [NP DirObj PP-with]
   [NP DirObj]
   [NP DirObj]
   [NP DirObj PP-with]
   [NP DirObj]
   [NP DirObj PP-for]
   ```

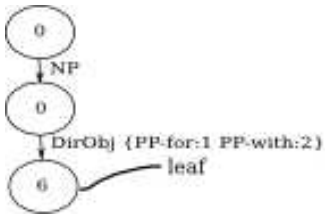
4. A trie is built with patterns:

   ```
   [NP DirObj] ->3
   [NP DirObj PP-with] ->2
   [NP DirObj PP-for] ->1
   ```

5. Leafs in the trie are “DirObj”,“PP-with” and “PP-for”. Since DirObj also occurs in the trie in a position other than leaf, it will not be considered as an adjunct in this iteration. In contrast, both PP-with and PP-for fulfill the conditions to be considered adjuncts, so we prune the patterns the trie, which will now have the single pattern, which forms a trie with 2 adjuncts (with information about the number of occurrences of each adjunct constituent):

   `[NP DirObj {PP-with:1 PP-for:2}]`

6. If the trie has been modified in this iteration, we go back to 2. If no modification has been operated, the procedure ends.

Moreover, the user can also specify a minimum number of occurrences of a verb to be taken into consideration, thus ruling out verbs for which there is not enough evidence in the corpus to obtain reliable subcategorization information.

7 Examples of applications

We have applied IRASubcat to two very different corpora in order to test its functionalities.

We have applied it to the SenSem corpus (Castellón et al., 2006), a corpus with 100 sentences for each of the 250 most frequent verbs of Spanish, manually annotated with information of verbal sense, syntactical function and semantic role of sentence constituents, among other information. From all the available information, we specified as input parameter for IRASubcat to consider only the syntactic function of sentence constituents. Thus, the expected output was the syntactic aspect of subcategorization frames of verbs. We worked with the verbal sense as the unit.

We compared the patterns associated to each verbal sense by IRASubcat with the subcategorization frames manually associated to the verbs at the a lexical data base of SenSem verbs[1]. We manually inspected the results for the 20 most frequent verbal senses. Results can be seen at Table 1. We found that the frequency threshold was the best filter to associate patterns and verbs, obtaining an f-measure of 74%. When hypothesis tests were used as a criterion to filter out associations of patterns with verbal senses, performance dropped, as can be seen in the lower rows of Table 1.

We also applied IRASubcat to an unannotated corpus of Russian. The corpus was automatically POS-tagged with TreeTagger (Schmid, 1994). We applied IRASubcat to work with parts of speech to build the patterns.

We manually inspected the patterns associated to prototypical intransitive ("*sleep*"), transitive ("*eat*") and ditransitive ("*give*") verbs. We found that patterns which were more strongly associated to verbs corresponded to their prototypical behaviour. For example, the patterns associated to the verb "*eat*" reflect the presence of a subject and a direct object:

Pattern	occurrences	% Likelihood Ratio Test
['V', 'Nn']	5	99
['V', 'C']	5	95
['V', 'R']	4	did not pass
['V', 'Nn', 'C', 'Q']	3	95
['V', 'V', 'Nn', 'Nn']	3	99
['V', 'Nn', 'Na']	3	99,5
['Nn', 'C']	3	90
['V', 'Nn', 'Nn']	3	99
['V', 'R', 'Q']	2	95
['V', 'Nn', 'An']	2	99

For more details on evaluation, see (Altamirano, 2009).

8 Conclusions and Future Work

We have presented a highly flexible tool to acquire verbal subcategorization information from corpus, independently of the language and level of annotation of the corpus. It is capable of identifying adjuncts and performs different tests to associate patterns with verbs. Thresholds for these tests can be set by the user, as well as a series of other system parameters. Moreover, the system is platform-independent and open-source[2].

We are currently carrying out experiments to assess the utility of the tool with two very different corpora: the SenSem corpus of Spanish, where sentences have been manually annotated with information about the category, function and role of the arguments of each verb, and also a raw corpus of Russian, for which only automatic part-of-speech tagging is available. Preliminary results indicate that, when parameters are properly set, IRASubcat is capable of identifying reliable subcategorization information in corpus.

As future work, we plan to integrate evaluation capabilities into the tool, so that it can provide precision and recall figures if a gold standard subcategorization lexicon is provided.

Acknowledgments

This research has been partially funded by projects KNOW, TIN2006-15049-C03-01 and *Representation of Semantic Knowledge* TIN2009-14715-C04-03 of the Spanish Ministry of Education and Cul-

[1]The lexical data base of SenSem verbs can be found at `http://grial.uab.es/adquisicio/`.

[2]IRASubcat is available for download at `http://www.irasubcat.com.ar`

applied filter	Precision	Recall	F-measure
Frequency	.79	.70	**.74**
likelihood ratio 90%	.42	.46	.44
likelihood ratio 95%	.38	.42	.40
likelihood ratio 99%	.31	.36	.335
likelihood ratio 99.5%	.25	.28	.265

Table 1: Performance of IRASubcat to acquire subcategorization information from the SenSem corpus, for the 20 most frequent verbal senses, as compared with manual association of subcategorization patterns with verbal senses. Performance with different filters is detailed: only the most frequent patterns are considered, or only patterns passing a hypothesis test are considered.

ture, and by project PAE-PICT-2007-02290, funded by the National Agency for the Promotion of Science and Technology in Argentina.

References

Bibliographisches Institut & F. A. Brockhaus AG, editor. 2001. *Duden das Stilwörterbuch*. Dudenverlag.

I. Romina Altamirano. 2009. Irasubcat: Un sistema para adquisición automática de marcos de subcategorización de piezas léxicas a partir de corpus. Master's thesis, Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Argentina.

Michael R. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Comput. Linguist.*, 19(2):243–262.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. pages 356–363.

J. Carroll and A. Fang. 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114.

J. Carroll, G. Minnen, and T. Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-conference Workshop on Linguistical ly Interpreted Corpora*, pages 35–41, Bergen, Norway.

Irene Castellón, Ana Fernández-Montraveta, Glòria Vázquez, Laura Alonso, and Joanan Capilla. 2006. The SENSEM corpus: a corpus annotated at the syntactic and semantic level. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Paula Chesley and Susanne Salmon-Alt. 2006. Automatic extraction of subcategorization frames for french.

Grzegorz Chrupala. 2003. Acquiring verb subcategorization from spanish corpora. Master's thesis, Universitat de Barcelona.

Erika de Lima. 2002. The automatic acquisition of lexical information from portuguese text corpora. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*.

Judith Eckle-Kohler. 1999. Linguistic knowledge for automatic lexicon acquisition from german text corpora.

Charles J. Fillmore. 1968. The case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, New York.

Effi Georgala. 2003. A statistical grammar model for modern greek: The context-free grammar.

Alexandra Kinyon and Carlos A. Prolo. 2002. Identifying verb arguments and their syntactic function in the penn treebank. pages 1982–1987.

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 199–206, Morristown, NJ, USA. Association for Computational Linguistics.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. pages 235–242.

Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. 2005. Large-scale induction and evaluation of lexical resources from the penn-ii and penn-iii treebanks. volume 31, pages 329–365.

Evan Sandhaus, editor. *New York Times*.

Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for czech. pages 691–697.

H. Schmid. 1994. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *COLING'00*, pages 747–753.

Kristina Spranger and Ulrich Heid. 2003. A dutch chunker as a basis for the extraction of linguistic knowledge.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate arguments structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*.

Akira Ushioda, David A. Evans, Ted Gibson, and Alex Waibel. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. pages 95–106.

Oliver Wauschkuhn. 1999. Automatische extraktion von verbvalenzen aus deutschen text korpora. Master's thesis, Universität Stuttgart.

WSJ, editor. 1994. *Wall Street Journal*.